# Türkçe Ders Metinleri İçin Özelleştirilmiş Bir Varlık İsmi Tanıma Yapısı

# A Named Entity Recognition Structure Specialized for Turkish Lecture Notes

Önder Can SARI[1], Özlem AKTAŞ[2]
[1]Dokuz Eylül Üniversitesi, Bilgisayar Mühendisliği Anabilim Dalı, İzmir, Türkiye,
onder.sari@ceng.deu.edu.tr
[2]Dokuz Eylül Üniversitesi, Bilgisayar Mühendisliği Bölümü, İzmir, Türkiye
ozlem@cs.deu.edu.tr

## Öz

*Varlık ismi tanıma (VİT); doğal dil işleme ve metin madenciliği alanlarının kapsamında yer alan bir bilgi çıkarımı görevidir. Kapsam ve kullanılan metotlar açısından, çalışmalar arasında farklılıklar görülse de temel olarak, bir metin içerisindeki kişi, yer, kurum-kuruluş vb. belirten ifadelerin doğru şekilde tespit edilmesini hedefler. Bu çalışmada, Türkçe yazılmış ders metinleri (tarih ve coğrafya alanlarında) için bir VİT yapısı geliştirilmiştir. Tek başına ele aldığımızda bu yapı, bir bilgi çıkarımı görevi doğrultusunda özelleştirilmiş bir projedir. Bunun yanı sıra çalışmanın eğitimsel bir değeri de vardır; çünkü sistemden beklenen sonuç, verilen ders metninin içeriğinden anlamlı kelime ya da kelime grupları bulunmasıdır ki; bu da farklı dersler ya da ders konuları için terimler sözlüğü yapıları oluşturmak için kullanılabilir. Oluşturulan sözlüklerin, bir ders metninin içeriğindeki soru değeri taşıyabilecek ifadelerin tespitine ve sınav hazırlama sürecine yardımcı olması hedeflenmektedir. Bu makalede, VİT görevi ve görevin kapsamı hakkında genel bilgi verilmiş; alanda yapılmış önceki çalışmalardan bahsedilmiş; bu çalışma doğrultusunda geliştirilen sistem tanıtılmış; sistemin başarısı, yapılan deney sonuçları üzerinden değerlendirilmiş ve geliştirme-iyileştirme olanakları hakkında yorumlar paylaşılmıştır.*

**Gönderme ve kabul tarihi:** 22.10.2018-08.11.2018
**Makale türü:** Araştırma

***Anahtar Kelimeler:*** *Bilişimsel dilbilim, varlık ismi tanıma, doğal dil işleme, bilgi çıkarımı, eğitimsel teknoloji*

## Abstract

*Named entity recognition (NER) is an information extraction (IE) task that is in the scope of natural language processing (NLP) and text mining. Its extent and methods may differ between studies, but basically, it aims to detect expressions that indicates a person, location, organization etc. In this study, a NER structure is developed for Turkish lecture notes (for history and geography courses). Separately, this structure is a project that is specialized for an IE task. Besides, it also has an educational value, as the projected outcome from its execution is meaningful words or word groups from the content of input lecture notes, which can be used to construct glossary of terms structures for individual courses or course subjects. With these glossary of terms structures, it is aimed to detect expressions in the content of a lecture note that can be used for questions and support a test preparation process. In this document, general information about NER task and its scope is given; previous studies on the field are mentioned; the system developed in line with this study is introduced; success of the system is evaluated through experiment results and some thoughts for enhancement are shared.*

***Keywords:*** *Computational linguistics, named entity recognition, natural language processing, information extraction, educational technology.*

## 1. Giriş

Varlık ismi (*named entity)*, bir özel isim kullanılarak atıfta bulunulan tüm varlıkları kapsayan bir ifadedir. Bir bilgi çıkarımı görevi olan VİTise, bir metin içerisindeki varlık isimlerinin tespit edilmesini ve önceden tanımlı kategoriler göz önüne alınarak sınıflandırılmasını hedefler. VİT birleşik bir görev olarak ele alınmalıdır; çünkü sırayla yerine getirmesi gereken iki gereksinim içerir: İlki, bir özel isim ifade eden metin parçalarının sınırlarını doğru belirlemek; ikincisi ise bu ifadeleri doğru kategoriler altında sınıflamaktır.

Haber metinleri üzerinde çalıştırılan genel yapılı VİT sistemleri, kişi, yer ve kurum isimlerinin tespitine yoğunlaşmıştır. Daha özelleşmiş uygulamalarda ise ticari ürün, sanat eseri, bazı biyolojik terim (protein, gen türleri gibi) isimleri gibi farklı kategoriler karşımıza çıkabilir [1]. Çizelge 1'de örnek varlık ismi türleri ve kapsamları gösterilmiştir. VİT sistemlerinin çoğunda, varlık ismi kavramının özel isimlerle sınırlı tutulmadığı, metin içerisindeki karakteristik anlama sahip ifadelerin (özel isim olmasa dahi) de bu kapsamda değerlendirildiği gözlenir. Bu durum tarih, saat gibi zamansal ifadelerin ya da ölçüm, sayım, fiyat gibi sayısal ifadelerin de varlık ismi kategorilerine (etiket olarak da adlandırılır) dahil edilmesine yol açabilir. Bu makalede detaylandırılan sistem, tarih ve coğrafya alanlarında yazılmış Türkçe ders metinleri üzerine özelleştirilmiştir. Tespit edilen karakteristik varlık isimlerinin, tarih ve coğrafya alanlarına özel bir terimler sözlüğü yapısının altyapısını oluşturması hedeflenmiştir.

**Çizelge 1. Varlık ismi türleri ve işaret ettikleri varlıklara örnekler**

| Tür | Etiket | Örnek Kategoriler |
|---|---|---|
| Kişi | PER | Şahıslar, hayali karakterler, küçük topluluklar |
| Kurum | ORG | Şirketler, acenteler, siyasi partiler, spor kulüpleri |
| Yer | LOC | Fiziki alanlar, dağlar, göller, denizler |
| Jeopolitik İfade | GPE | Ülkeler, eyaletler, şehirler, ilçeler |
| Tesis | FAC | Köprüler, havayolları, binalar |
| Taşıt | VEH | Uçaklar, trenler, arabalar |

VİT sistemlerinde genel işleyiş, girdi olarak işaretlenmemiş (*unannotated*) bir metin bloğu alıp, çıktı olarak tespit edilen varlık isimlerini gösteren, yani işaretlenmiş (*annotated*) bir metin bloğu döndürmektir. Örneğin işaretlenmemiş "Mustafa Kemal Atatürk 1881 yılında Selanik'te doğdu." cümlesi için beklenen çıktı "[Mustafa Kemal Atatürk]$_{\text{PER}}$ [1881]$_{\text{DATE}}$ yılında [Selanik]$_{\text{LOC}}$'te doğdu." şeklindedir.

Çok anlamlılık ya da anlam bulanıklığı (*ambiguation*), birçok doğal dil işleme görevinde olduğu gibi, VİT sistemleri için de önemli bir sorundur. Örneğin "Washington" kelimesi, bulunduğu içeriğe göre bir kişi, bir yer, bir kurum (spor kulübü) ya da bir taşıt (gemi) ismi belirtiyor olabilir. Ya da "Ural" kelimesi bir yer (nehir ismi) ya da kişiden bahsediyor olabilir. VİT sistemleri, başarı oranlarının buna benzer durumlardan olumsuz etkilenmesini önlemek adına farklı yaklaşımlardan faydalanabilir.

Sözcük birimleştirme (*tokenization*), kelime düzeyinde yapılan bir metin bölümlendirme operasyonudur ve VİT işlemi için yaygın bir başlama noktasıdır. İstatistiksel metotlar tercih ediliyorsa, sonrasında sekans etiketleme (*sequence labeling*) ile devam edilir. Bu işlemde, sınıflandırıcılar (*classifier*), sözcük birimleri (*token*) sistemde tanımlı belirli bir türe ait bir varlık ismi işaret edip etmediğine göre etiketlemek üzerine eğitilirler. Bir varlık isminin başlangıcını (B →*Beginning*) ya da devamını (I →*Inside*) içeren, ya da hiç varlık ismi içermeyen (O →*Outside*) kelimeleri ayrıştırmayı hedefleyen IOB, etiketleme için sık tercih edilen bir formattır. Çizelge 2'de, sözcük birimlerin etiketlenmesi için kullanılan IO ve IOB formatları arasındaki farklar, bir cümle üzerinde gösterilmiştir.

**Çizelge 2. IO ve IOB kodlama farkına örnek bir sekans etiketleme**

| | IO Kodlama | IOB Kodlama |
|---|---|---|
| Mehmet | PER | B-PER |
| Edvard | PER | B-PER |
| Munch | PER | I-PER |
| 'un | O | O |
| resmini | O | O |
| Ahmet | PER | B-PER |
| 'e | O | O |
| gösterdi | O | O |

Öznitelik seçimi (*feature selection*), VİT başarısını etkileyebilecek başka bir faktördür. Sözcük birimler ile ilgili elde edilen bulguların tutulduğu öznitelikler, sistemin daha isabetli tahminler yapması için

kullanılan yapılardır. Sistem eğitilirken birden fazla öznitelik (*feature*) kullanılabilir. Örneğin yaygın özniteliklerden biri olan şekil (*shape*), sözcük birimlerin yazımı hakkında karakter düzeyinde bilgi tutar (tamamı küçük harf, tamamı büyük harf, sadece ilk harf büyük, tire karakteri içeriyor vb.). Dilbilgisi kurallarına uygun yazılmış metinlerde, büyük harf kullanımı ve noktalama işaretleri, varlık isimlerini tespit etmek için önemli ipuçları verebilirler. Çoğu VİT sistemi, yer adları dizini (*gazetteer*) yapılarından faydalanır. Kurum-kuruluş isimleri ve biyolojik terimler için de benzer yapılar mevcuttur. Saygı ifadeleri ya da unvanları barındıran, kestirimci kelime (*predictive words)* listeleri de kullanılabilir. Liste kullanımı ile, sözcük birimlerin ilgili dizin ya da listede bulunup bulunmadığı bilgisini içeren öznitelikler elde edilebilir. Sözcük türü etiketi (*part-of-speech tag),* kök bulma (*stemming)* sonrası ifade, söz öbeği etiketi (*syntactic chunk label)* diğer yararlanılabilir özniteliklerden birkaçıdır.

VİT algoritmaları temel olarak üç kategoride incelenir: İstatistiksel, kural tabanlı ve karma (*hybrid)* modeller. İstatistiksel modellerdeki temel yaklaşım, varlık isimlerine ait kuralların ve örüntülerin, önceden işaretlenmiş eğitim verisi (*training data)* yardımıyla öğrenilmesidir. Eğitim verisi, eğer varsa kullanılacak öznitelikler hakkında da bilgi sağlayacak şekilde etiketlenmelidir. Saklı markov modelleri (*hidden markov models),* maksimum entropi ve koşullu rassal alanlar (*conditional random fields)*, en yaygın istatistiksel modellerdir. Kural tabanlı modeller, öznitelik kümelerinden elde edilen bulguların, kullanılan dil ile ilgili önceden tanımlanmış dilbilgisi ve örüntü kurallarına göre değerlendirilmesi esasına dayanır.

Günümüzde başarı oranı yüksek bir VİT sistemi, haber içeriklerinin sınıflandırılması, tavsiye sistemleri, müşteri destek sistemleri, sosyal medya analizi, duygu analizi, istenmeyen (*spam)* e-posta tespiti, literatür taraması, ya da bu çalışmada önerilen modelin geliştirilme sebebi olan eğitimsel amaçlar gibi birçok farklı kullanım senaryosunda fayda sağlayacaktır.

## 2. İlgili Bilimsel Çalışmalar

Mesaj Anlama Konferansları (*MUC – Message Understanding Conferences)* ile bilgi çıkarımı çalışmaları teşvik edilerek alanda ilerleme sağlamak hedeflenmiştir. Başlıca iki değerlendirme kriterikesinlik (*precision)* ve hassasiyet (*recall),* MUC-2'de detaylandırılmış ve verilen bilgi çıkarımı görevlerinde kullanılmıştır. 1996'da düzenlenen altıncı konferansta İngilizce için VİT, verilen görevlerden biri olmuş ve en başarılı sistemde %97 kesinlik ve %96 hassasiyet değerlerine ulaşılmıştır. Eğitim verisi olarak Wall Street Journal makaleleri işaretlenmiştir. ENAMEX (kişi, kurum ve yer adları için) ve NUMEX (saat, para ifadesi, yüzde için) etiketleri bu konferansta tanıtılmıştır [2].

Cucerzan ve Yarowksy (1999) [3], Türkçe üzerine yayımlanmış ilk VİT araştırmasıdır. Sistem dillerden bağımsız olup, karakter düzeyinde oluşturulan bir ağaç yapısı üzerinde yinelemeli öğrenme (*iterative learning)* temelli bir önyükleme (*bootstrapping)* algoritmasından faydalanır. Sistem eğitimi bir kelimenin, bir belge içinde genellikle belirli tek bir anlam ifade ettiği ön kabulü üzerinden ilerler. Kaynak dil ile ilgili küçük bir varlık ismi listesi ile süreç başlatılıp, metinlerden morfolojik ve bağlamsal (*contextual)* ipuçları elde etmeye çalışılır. Örneğin, "-escu" ifadesi Rumence için neredeyse hatasız bir soy ismi göstergesi olarak bulunmuştur. Sistemin Türkçe için başarısı %60 kesinlik, %47 hassasiyet ve %53 f-ölçütü olarak hesaplanmıştır.

Alfonseca ve Manandhar (2002) [4], WordNet ontolojisinden yararlanarak, sınıflandırılmamış bir kavram için en isabetli kapsayıcı (*hypernym)*terimi bulmayı hedeflemişlerdir. Sistem sınıflandırma için, arama motorları üzerinde çalıştırdığı sorgular ile, aday kelimeler için benzerlik skorları elde ederek çalışır. Buradaki yaklaşım, anlamsal (*semantic)* olarak birbiriyle bağlantılı kelimelerin bir arada bulunmasının muhtemel olduğudur.

Tür vd. (2003) [5], n-gram dil modellerini saklı markov modelleri içine entegre ederek bir yapı kurmuşlardır. Çalışma dört farklı model içerir: Sözcüksel (*lexical*) modelde, sözcük birimler arası boşluklar, varlık isimlerinin sınırlarını ifade eden *yes, no* ve *mid* işaretleri (*boundary flag)* ile etiketlenir. Bağlamsal model, varlık ismi türü bilinmeyen bir sözcük birimin (*unk* etiketli), metin içerisinde önünde ve arkasında yer alan diğer sözcük birimler yardımıyla sınıflandırmayı hedefler. Çizelge 3'te bağlamsal modelin bilinmeyen bir kelime için kullanımına örnek verilmiştir. Morfolojik model, sözcük birimler karakter düzeyinde ele alır (tamamı büyük harf, sadece ilk harfi büyük gibi); ayrıca Türkçe için bir kişi, yer ve kurum isimleri sözlüğünden yararlanır. Etiket modeli (*tag model)* ise varlık ismi türü belirten

etiketler ile (kişi, yer, kurum, diğer) sınır belirten etiketler (yes, no, mid) arasındaki üçlükombinasyonların (*trigram)* olasılıkları ile ilgilenir. Deneyler için gazete makaleleri kullanılmıştır. Tüm modeller birleştirildiğinde varlık ismi metinleri için %90.4, varlık ismi türleri için %92.7 doğruluk (*accuracy)* elde edilmiştir.

**Çizelge 3. Bağlamsal modelin bilinmeyen bir kelime için kullanımı (Tür vd. [2003])**

| Trigram Sekans | Olasılık |
|---|---|
| Dr./*diğer* boşluk/*yes* unk/*kişi* | 0.990119 |
| Dr./*diğer* boşluk/*yes* unk/*yer* | 0.000690 |
| Dr./*diğer* boşluk/*yes* unk/*kurum* | 0.000880 |
| Dr./*diğer* boşluk/*yes* unk/*diğer* | 0.002688 |

CoNLL konferansları katılımcılara bilişimsel dilbilim görevleri veren başka bir etkinliktir. 2003'teki konferans dillerden bağımsız bir VİT yapısı oluşturma üzerinedir. Katılımcılardan ek görev olarak da işaretlenmemiş veriyi eğitim sürecine katmaları istenmiştir. Bu veriden büyük harf kullanımı ile ilgili bilgi edinmenin sonuçlara etkisinin, verinin yeni dizin terimleri bulmak için kullanılmasından daha olumlu olduğu gözlemlenmiştir [6].

Wentland vd. (2008) [7], HeiNER isimli çokdilli bir varlık ismi kaynağı oluşturmuştur. Wikipedia, varlık isimlerini elde etmek için temel kaynak olarak kullanılmıştır. Kelimelere ait belirsizlikleri giderecek bir sözlük oluşturmak için, Wikipedia'daki adlandırma (*disambiguation)* ve yönlendirme (*redirect)* sayfalarından faydalanılmıştır. Wikipedia makale başlıklarının bir varlık ismi belirtme olasılığının yüksek olduğu; bu durumun morfolojik normalleştirme ya da varlık ismi sınırlarının tespiti gibi bazı görevleri kolaylaştırdığı gözlemlenmiştir.
Küçük ve Yazıcı (2009a) [8], Türkçe için kural tabanlı bir VİT sistemi oluşturup, başarısını farklı alanlardaki (haber metinleri, masallar, tarihi metinler) metinlerde test etmişlerdir. Sistem Türkçe kişi isimleri, tanınmış siyasetçiler, iyi bilinen kurum-kuruluş isimleri ve muhtemel örüntüler gibi farklı sözlüksel kaynaklardan faydalanmıştır. Sonuçlara göre haber metinlerinde görülen %78 f-ölçütü, alan değişiminden olumsuz etkilenerek masallarda %69, tarihi metinlerde %55 olarak ölçülmüştür. Yabancı kişi isimleri ve tarihi kişi-kurum isimlerinin sözlüksel kaynaklarda olmamasının bu performans düşüşünün temel nedenlerinden olduğu gözlemlenmiştir.

Küçük ve Yazıcı (2009b) [9], çalışmalarını videolardan elde edilen metinler üzerinde de test etmişlerdir. Bunun için TRT arşivinden seçilen 16 haber videosu metne dökülmüştür. Çalışmanın yapıldığı dönemde Türkçe için bir konuşma tanıma (*speech recognition*) sistemi olmadığı için bu işlem elle yapılmıştır. Başarı %73 kesinlik, %77 hassasiyet ve %75 f-ölçütü olarak ölçülmüştür.

Tatar ve Çiçekli (2011) [10], VİT sistemlerinin alan değişimlerinden olumsuz etkilenmesini önlemeyi amaçlayarak, gözetimli öğrenme (*supervised learning)* ile otomatik kural tanımlama üzerine çalışmışlardır. Sistem yazımsal (*orthographical)*, bağlamsal, sözcüksel ve morfolojik özniteliklerden yararlanmış, ayrıca 2 seviyeli sözlüksel kaynaklar kullanmıştır. Türkçe haber metinleri içeren TurkIE veri kümesinde yapılan testlerde %91.7 kesinlik, %90 hassasiyet ve %91 f-ölçütü değerlerine ulaşılmıştır.

Küçük ve Yazıcı (2012) [11], kural tabanlı sistemleri üzerinden devam ederek karma bir model kurmuşlardır. n (bir kelimenin görülme sayısı) ve p (aynı kelimenin varlık ismi olarak işaretlenmiş şekilde görülme sayısı) şeklinde iki istatistiksel öznitelik tanımlayarak; p/n değerini o kelime için güven değeri olarak kullanmışlardır. Sistem üç farklı alanda yeniden test edildiğinde haber metinleri için %85.9, masallar için %85 ve tarihi metinler için %66.9 f-ölçütü değerlerine ulaşıldığı görülmüştür.

Şeker ve Eryiğit (2012) [12], koşullu rassal alanlar (CRF) ile çalışarak bir istatistiksel model kurmuşlardır. CRF modelinde öznitelikler için pencere genişliği {-3,+3} şeklinde tanımlanmıştır. Dizinler kullanılmış; ilaveten normal kelimelerden sonra veya önce gelerek varlık ismi oluşturabilecek üretici kelimeler listesi (22 kişi, 44 yer, 60 kurum ismi için) de kullanılmıştır. Üç kategoriye ayrılmış (morfolojik, sözcüksel, dizin tarama) toplam 14 öznitelik tanımlanmış; bunlar sisteme birer birer eklenerek başarıya katkıları ölçülmüştür. Deney sonuçları cümle başını belirten SS (*start of sentence*) dışında tüm özniteliklerin sistem başarısını olumlu etkilediğini göstermiştir. Sistem son durumda MUC kriterlerine göre %94.6, CoNLL kriterlerine göre %91.9 f-ölçütü değerlerine erişmiştir.

Küçük vd. (2014) [13], Türkçe üretilmiş Twitter gönderileri üzerinde VİT deneyleri gerçekleştirmiştir. Klasik kategorilerden (çalışmada PLO olarak isimlendirilmiş) ayrı olarak bir "*misc*" (ticari ürün, televizyon programı, müzik grubu isimleri gibi) türü

eklenmiştir. Kişi isimleri için aranan isim – soy isim ikilisi şeklinde olma şartının ihmal edildiği durumlara sık rastlandığı için; Avrupa Basın Takip (*EMM: Europe Media Monitor)* veri tabanı taranarak, tek kelime şeklinde sık kullanılan (En az 30 kere geçme şartı aranmıştır) kişi ve kurum isimleri bulunmuş ve iki liste elde edilmiştir. Deney sonuçları, veri setindeki PLO kullanımlarında %25 oranında büyük harf kullanımı hatası olduğunu, kişi isimlerinin isim – soy isim ikilisi şeklinde yer alma oranının sadece %32 olduğunu ve PLO metinlerinde %10 oranında Türkçe karakter problemi olduğunu göstermiştir. Konu etiketi (*hashtag*) metninde bulunan, birden çok kelimeden oluşan varlık isimlerinin tespiti de boşluk kullanımı olmadığı için başka bir sorun olarak ortaya konmuştur. Sistem %66 kesinlik, %31.5 hassasiyet ve %42.6 f-ölçütü değerlerine ulaşmıştır.

Küçük ve Arıcı (2016) [14], ODTÜ Türkçe Derlem içerisinden seçtikleri 10 gazete makalesi üzerinde varlık isimlerini işaretleyerek, Türkçe için 1425 varlık ismi (398 kişi, 567 yer, 460 kurum ismi) içeren bir veri seti ortaya çıkarmışlardır.

Şeker ve Eryiğit (2016) [15], önceki çalışmaları üzerinden ilerleyerek, kullanıcı tarafından oluşturulmuş içerik (*UGC: User generated content*) üzerinde çalışmışlardır. Büyük harf kullanımının eksikliği kaynaklı performans düşüklüğünü önlemek için, çok anlamlı kelimeler arasından cins isim olarak kullanılma ihtimalleri düşük olanları belirleyerek "ilk harfi otomatik olarak büyük harfe çevirme listesi" (*CAP: Auto capitalization gazetteer)* elde etmişlerdir. Önceki çalışmanın aksine, öznitelikler sistemden birer birer çıkararak sisteme katkıları ölçülmüştür. SS özniteliğinin, bu ölçümde sisteme %2.11 olumlu katkı verdiği görülmüştür. UGC veri setindeki deneylerde %67.9 başarıya ulaşılmıştır. Ayrıca CAP özniteliğinin sistemden çıkarılmasının, %20 üzeri bir performans kaybına yol açtığı gözlemlenmiştir.

Ertopçu vd. (2017) [16], parametrelerin değiştirildiği farklı metotlar ile testler yaparak, en yüksek başarı oranını sağlayan parametrelere ulaşmaya çalışmışlardır. En başarılı sonuçların, sınıflandırma algoritması olarak çok katmanlı algılayıcı (*multilayer perceptron)* seçilip öğrenme hızı *(learning rate)*0.1, pencere genişliği 1 verildiğinde ve 7 öznitelik kullanıldığında (büyük harf kontrolü, tarih ifadesi kontrolü, saat ifadesi kontrolü, kesir ifadesi kontrolü, sözcük türü, kök formu, gövde formu) elde edildiği (%7.64 hata oranı ile) gözlemlenmiştir.

Şahin vd. (2017) [17], Türkçe Vikipedi sayfalarının otomatik olarak sınıflandırılmasıyla VİT için bir Türkçe derlem oluşturmuşlardır. Derlem 300 bine yakın terim içermektedir. Terimler 4 ana kategoriye (kişi, yer, kurum, diğer) bağlı toplam 77 alt kategoriyle ilişkilendirilmiştir. Kelime belirsizliklerinin önüne geçmek için Freebase bilgi tabanından yararlanılmıştır. Sistem VİT için %84 f-ölçütü değerlerine ulaşmıştır.

Güneş ve Tantuğ (2018) [18], iki yönlü uzun-kısa süreli bellek (*bidirectional long-short term memory*) yapay sinir ağı yapısını 5 farklı modelde kullanmışlardır. CoNLL ölçümleri baz alınarak yapılan deneylerde, temel veri kümesi için %91.59 f-ölçütü değerine ulaşılmıştır. Sözcük vektörlerini kullanan temel veri kümesinin, yazım özellikleri ve biçimbilim özellikleriyle desteklendiği katmanlı bir yapay sinir ağı modeli önerilmiştir. Son modelin sistem başarısı %93.69 seviyesine yükselmiştir.

Güngör vd. (2018) [19] önerdikleri yapay sinir ağı modelinde, cümledeki sözcükleri baştan ve sondan işleyerek konum bilgisinin tutulduğu karakter tabanlı sözcük vektörleri oluşturmuşlardır. Bu vektörlerin, dağılımsal sözcük vektörleri ile yakalanamayan sözcük içi ilişkilerini yakalamak için kullanılması hedeflenmiştir. Sadece dağılımsal sözcük vektörleri kullanıldığında %90.96 f-ölçütü olarak hesaplanan sistem başarısının, karakter tabanlı sözcük vektörlerinin eklenmesiyle %93.37 seviyesine yükseldiği gözlemlenmiştir.

## 3. Uygulama

Bu çalışmada önerilen VİT sistemi, eğitimsel amaçlar doğrultusunda geliştirilmiş bir bilgi çıkarımı yazılımıdır. Tarih ve coğrafya alanlarında yazılmış Türkçe ders metinleri için özelleştirilmiştir. Çalışmanın birincil hedefi, sisteme girdi olarak verilen metin belgelerinin içeriğindeki varlık isimlerinin yüksek başarıyla tespit edilmesi olup; bu varlık isimleri ile nitelikli ve beklentilere cevap veren terimler sözlüğü yapılarının temelini oluşturmak bir sonraki hedef olarak belirlenmiştir. Araştırmanın uzun vadeli hedefi ise, oluşturulacak terimler sözlüklerine, sınav hazırlama süreçlerine destek sağlayabilecek bir yapı kazandırmaktır.

## 3.1. Önerilen Sistem

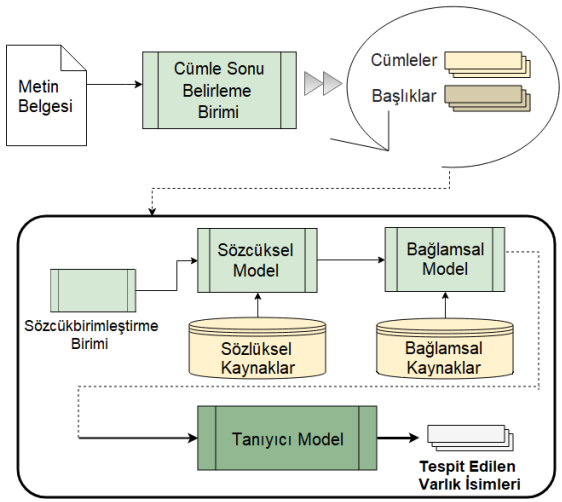
Önerilen VİT sistemi, kural tabanlı bir model kullanılarak inşa edilmiştir. Girdi olarak bir metin belgesi alıp, tespit edilen varlık isimlerini ve türlerini çıktı olarak vermektedir.

Sistem cümleler üzerinde çalışacak şekilde geliştirilmiştir. Bu nedenle girdi olarak alınan metin belgesi ilk olarak cümle sonu belirleme (*SBD*: *sentence boundary detection*) birimine gönderilir. Bu birim, aldığı metin belgesini ilk olarak bir ön işleme (*pre-processing)* sürecinden geçirir. Bu süreç gereksiz karakter, sembol ve boşlukların çıkarılması, maddelerle ayrılmış metin parçalarının birleştirilmesi gibi operasyonları kapsamaktadır. Ön işlemeden sonra başlıklar ve cümle sonları tespit edilir ve kullanıcıya cümleler ve başlıklar olmak üzere 2 sıralı liste döndürülür. Bu işlemler için cümle sonu koşullarının tanımlandığı bir liste oluşturulmuş; sonrasında bu koşullar arka uçta kurallı ifadelere (*regular expression*) dönüştürülerek SBD biriminin kullanımına sunulmuştur. Metin içerisindeki kısaltma kullanımlarının yol açabileceği hatalı sonuçların önüne geçebilmek adına, bu aşamada 204 elemandan oluşan bir Türkçe kısaltma listesi kullanılarak bir kısaltma denetim operasyonu da gerçekleştirilir.
Çizelge 4'te, SBD birimi çalıştırılırken kullanılan cümle sonu koşullarından bazıları gösterilmiştir (KH → küçük harf, BH → büyük harf, B → boşluk karakteri, R → rakam; Doğru → cümle sonu koşulunu işaret eder, Yanlış → cümle sonu olmama koşulunu işaret eder).

**Çizelge 4. Örnek cümle sonu koşulları**

| Koşul | Çıktı |
|---|---|
| KH . BH | Doğru |
| KH . KH | Yanlış |
| KH . R | Doğru |
| BH. KH | Yanlış |
| KH .B . BH | Doğru |
| KH .B . R | Doğru |

SBD biriminin çalıştırılmasının ardından, metinden cümleler elde edilerek VİT sisteminin kullanımına sunulur. Dolayısıyla, VİT modelinin başarısı, SBD biriminin başarısına da bağlıdır. Cümleler VİT yapısına birer birer verilir ve sırasıyla sözcük birimleştirme birimi (*tokenizer*), sözcüksel model ve bağlamsal modelde işlenir. Bu üç birim, verilen cümleyi etiketleyerek, son birim olan tanıyıcı modele (*recogniser model)* bilgi sağlar. Tanıyıcı model, varlık isimlerini tespit etmek için, etiketlenmiş sözcük birimlerden oluşan cümleyi analiz eder. Şekil 1'de, önerilen sistem çatısına dair bir akış şemasına yer verilmiştir.




**Şekil 1.** VİT yapısı için önerilen sistem çatısı

## 3.2. Sözcük birimleştirme Birimi ve Sözcük birimler

Sisteme verilen metin belgesinden elde edilen cümleler ilk olarak sözcük birimleştirme biriminde işlenir. Bu birim, girdi cümleyi tarayarak kelime sonlarını ve noktalama işaretlerini tespit ederek cümleyi bir sözcük birim listesine dönüştürür. Sözcük birim yapılarının kapsamı, bir kelime ile sınırlı olmayabilir; noktalama işareti ya da noktalama işareti sonrası bir morfem de ifade edebilir. Dolayısıyla bu birimde yapılan işlemi, cümleyi kelimelerine ayırmak şeklinde ifade etmek hatalı olacaktır.

Program tarafında sözcük birimler, Token sınıfı nesnelerine atanır. Bu sınıfa ait nesneler, kendisinden önceki ve sonraki Token nesnesinin bilgisini de tutar. Bu tasarım, cümleden elde edilen sözcük birimlerin çift yönlü bağlı liste (*double linked list)* yapısında tutulmasını sağlamış olur. Token sınıfı nesneleri aynı zamanda, etiketleme işlemleri ile değer atamalarının yapılacağı öznitelik bilgilerini taşıyan bir dizi mantıksal değişken de içermektedir. Bir sözcük birimin etiketlenmesi ile, varlık isimlerinin tespiti

sırasında kullanılabilecek önemli bilgilerin sağlanması hedeflenmiştir.

Sözcük birimler üzerinde ilk etiketleme sözcük birimleştirme biriminde yapılır. Bu etapta yapılan etiketleme ile sisteme yazımsal, numerik, noktalama ile ilgili ve sözcük birim pozisyonu ile ilgili bilgiler sağlanmış olur. Çizelge 5'te, bu birimde kullanılan etiketler gösterilmiştir.

**Çizelge 5. Sözcük birimleştirme birimi etiketleri**

| Yazımsal Bilgi | Numerik Bilgi | Noktalama Bilgisi | Pozisyon Bilgisi |
|---|---|---|---|
| SW_CAPITAL | NUM | PUNCT_APOSTR | BEFORE_APOST |
| ALL_CAPITAL | ROMAN_NUM | PUNCT_OTHER_MID | AFTER_APOST |
| EW_DOT | ORD_NUM | PUNCT_OTHER_END | |
| | DAY_NUM | PERCT | |
| | MONTH_NUM | | |
| | YEAR_NUM | | |

- SW_CAPITAL: Sözcük birim metninin büyük harf ile başlayıp başlamadığı bilgisini tutar.
- ALL_CAPITAL: Sözcük birim metninin tamamen büyük harflerden oluşup oluşmadığı bilgisini tutar.
- EW_DOT: Sözcük birim metninin son karakterinin nokta olup olmadığı bilgisini tutar.
- NUM: Doğru (*true*) değeri atanması, sözcük birim metninin sayısal bir ifade belirttiğini gösterir.
- ROMAN_NUM: Doğru değeri atanması, sözcük birim metninin bir Romen rakamı belirttiğini gösterir.
- ORD_NUM: Doğru değeri atanması, sözcük birim metninin bir sıra sayısı belirttiğini gösterir.
- DAY_NUM: Doğru değeri atanması, sözcük birimin [1,31] aralığında sayısal bir ifade belirttiğini gösterir.
- MONTH_NUM: Doğru değeri atanması, sözcük birimin [1,12] aralığında sayısal bir ifade belirttiğini gösterir.
- YEAR_NUM: Doğru değeri atanması, sözcük birimin [100, 5500] aralığında sayısal bir ifade belirttiğini gösterir.
- PUNCT_APOSTR: Sözcük birim metninin bir kesme işareti olup olmadığı bilgisini tutar.
- PUNCT_OTHER_MID: Sözcük birim metninin, cümle ortasında kullanılan bir noktalama işareti (virgül, noktalı virgül, parantez vb.) olup olmadığı bilgisini tutar.
- PUNCT_OTHER_END: Sözcük birim metninin, cümle sonunda kullanılan bir noktalama işareti (nokta hariç) olup olmadığı bilgisini tutar.
- PERCT: Sözcük birim metninin yüzde işareti olup olmadığı bilgisini tutar.
- BEFORE_APOST: Doğru değeri atanması, bir sonraki sözcük birimin bir kesme işareti olduğunu gösterir.
- AFTER_APOST: Doğru değeri atanması, bir önceki sözcük birimin bir kesme işareti olduğunu gösterir.

## 3.3. Sözcüksel Model Kaynakları

Sözcüksel ve bağlamsal modeller, sözlüksel kaynaklar (*lexicon*) yardımıyla sözcük birimlerin etiketlenmesi ve son hâllerinin verilmesini amaçlar. Bağlaçların tutulduğu yardımcı liste haricinde, sözcüksel modelin kullandığı kaynaklar muhtemel özel isimlerin (kişi, yer ya da bölge belirtmek üzere) tutulduğu yapılardır.

- **TR_FirstNames:** Türkçe kişi isimlerinin tutulduğu kaynak listedir. TDK Kişi Adları Sözlüğü elemanlarını içeren bir veri tabanı baz alınarak hazırlanmıştır. Listenin ilk hâli 9699 eleman içerirken, sayı 3.3.1 alt başlığı altında detaylandırılacak elemelerden sonra 9619 elemana düşürülmüştür.
- **TR_CommonSurnames:** Sık karşılaşılan Türkçe soy isimlerinin tutulduğu kaynak listedir. Elemanlar, Wikipedia üzerindeki Türk sinema oyuncuları, Türk siyasetçiler (20. ve 21. yüzyıl), Türk yazarlar ve Türk Kurtuluş Savaşı'na katılan üst düzey subaylar listelerinden elde edilmiştir. Aynı kişinin isminin birden fazla yer aldığı durumlar (örneğin hem 20., hem 21. yüzyılda görev almış siyasetçiler) ve sık karşılaşılan soy isimlerinin tekrar ettiği durumlar elendiğinde, listenin son hâli 3039 eleman barındırmaktadır.
- **FRGN_FirstNames:** Sık karşılaşılan yabancı kişi isimlerinin tutulduğu kaynak listedir. Elemanlar, *ranker.com* üzerindeki "gelmiş

geçmiş en etkili insanlar" (*the most influential people of all time*) listesinden elde edilmiştir. Bu listede farklı ülkelerden, aralarında bilim adamlarının, politikacıların, sanatçıların, sporcuların, filozofların bulunduğu 2762 kişi bulunmaktadır. Liste verisi XML dosyası şeklinde çekilerek, bir normalleştirme işlemine tabi tutulmuş; sonucunda isimler, soy isimler ve ikinci isimler (göbek adları) elde edilmiştir. Normalleştirme fazı, İngilizce yazılmış kişi isimlerinde karşımıza çıkabilen "*of, the*" gibi tanımlık (*article*) ifadelerinin, sıra sayılarının, Romen rakamlarının ve unvan ya da lakap belirten (Aziz, Deli, Kral, Kraliçe, Baron, Prens, Prenses gibi) ifadelerin elenmesini kapsar. Tekrarlanan isimler de aynı şekilde listeden çıkarılır. Son durumda listede 1489 eleman bulunmaktadır.

- **FRGN_CommonSurnames:** Sık karşılaşılan yabancı soy isimlerinin tutulduğu kaynak listedir. Elemanlar yine *ranker.com* üzerindeki kaynak listeden elde edilmiştir. Son durumda listede 1864 eleman bulunmaktadır.
- **FRGN_MidNames:** Yabancı kişi adları ve soyadları arasında karşılaşılabilen "de, von, bin" gibi ifadelerin ya da bir büyük harf ve noktadan oluşan kısaltılmış ikincil adların tutulduğu kaynak listedir. Elemanlar yine *ranker.com* üzerindeki kaynak listeden elde edilmiştir. Son durumda listede 34 eleman bulunmaktadır.
- **Countries:** Ülke isimlerinin tutulduğu kaynak listedir. Günümüzde Birleşmiş Milletler üyesi olan 193 ülkenin, üye ülkelerin kapsamında değerlendirilen ülkelerin (İngiltere, Galler, İskoçya gibi) ya da onlara bağlı özerk ülkelerin (Porto Riko, Virjin Adaları) isimleri listeye eklenmiştir. Filistin, Tayvan ve KKTC de listeye eklenen diğer ülke isimleri olmuştur. İlaveten, tarihi metinlerde geçme ihtimali bulunan bazı yakın dönem ülkelerinin (Yugoslavya, SSCB gibi) isimleri de dahil edilmiştir. Son durumda listede 257 eleman bulunmaktadır.
- **TR_Cities:** Türkiye'nin 81 şehrinin isminin ve yaygın kullanılan farklı isimlendirmelerinin (Afyonkarahisar için Afyon gibi) tutulduğu kaynak listedir. Liste 86 eleman içermektedir.
- **TR_Districts:** Türkiye'nin ilçelerinin isimlerini tutan kaynak listedir. Listenin ilk hâli 984 elemandan oluşmaktayken, aynı ismi taşıyan ilçelerin ve bulunduğu şehrin ismiyle anılan merkez ilçelerin elenmesiyle beraber, son durumda listede 897 eleman bulunmaktadır.
- **FRGN_StatesCities:** Tüm ülkelerin başkentleri başta olmak üzere, yüksek nüfuslu ya da tarihi-turistik önemi yüksek dünya şehirlerinin isimlerinin tutulduğu kaynak listedir. Listeden ülkesiyle aynı adı taşıyan şehirler (Tunus, Cezayir, Singapur gibi) çıkarıldığında, son durumda listede 380 eleman bulunmaktadır.
- **GeographicRegions:** Kıta ve önemli coğrafi bölgelerin isimlerinin tutulduğu kaynak listedir. Liste 22 elemandan oluşmaktadır.
- **Conjunctions:** Türkçe'de kullanılan bağlaçların tutulduğu yardımcı listedir. Bu liste, cümle başında bulunduğu için büyük harfle başlayan bağlaçların tespit edilerek hatalı varlık ismi sonuçlarının önlenmesi açısından önemlidir.

### 3.3.1. Kaynaklardan Çıkarılan Elemanlar

Yabancı kişi isimleri elde etmek için kullanılan *ranker.com* üzerinden alınan listede Mustafa Kemal Atatürk, Orhan Veli Kanık, Yunus Emre, Halide Edip Adıvar gibi Türk kişiler de bulunuyor. Bu durumun da etkisiyle, TR_FirstNames ve FRGN_FirstNames listeleri arasında 29, TR_CommonSurnames ile FRGN_CommonSurnames listeleri arasında 13 ortak eleman olduğu gözlemlendi. Bu durum göz önüne alınarak listeler üzerinde güncellemeler yapıldı:

- "Abdullah, Selma, Selman, Zakir" gibi kelimeler iki listede de bırakıldı.
- "Edip, Evliya, Halide, Hamdi, Kemal, Mustafa, Orhan, Yunus, Ziya" gibi kelimeler FRGN_FirstNames listesinden çıkarıldı.
- "Adam, Alan, Boy, Sun, San" gibi kelimeler TR_FirstNames listesinden çıkarıldı.
- "Adıvar, Çelebi, Emre, Kanık, Pamuk, Atatürk, Tanpınar" gibi Türkçe soyadı belirten kelimeler FRGN_CommonSurnames listesinden çıkarıldı.
- "Bradley, Reynaud, Spence" gibi kelimeler TR_CommonSurnames listesinden çıkarıldı. Bu kelimelerin, yabancı kökenli olan ya da yabancı biriyle evli kişiler dolayısıyla ilk etapta bu listeye girdikleri görüldü.
- Bağlamsal model kaynaklarında bulunan bazı elemanlarla da kesişim görüldü. Bu elemanların

bir bölümü sözlüksel kaynaklardan çıkarıldı ve listelere son durumları verilmiş oldu.

## 3.4. Bağlamsal Model Kaynakları

Bağlamsal model tarafından kullanılan kaynaklar, metin içerisinde özel isimlere komşu olması muhtemel ifadelerin tutulduğu listelerdir. Bu ifadelerin varlık ismi metnine dahil olup olmamasına dair kesin bir kural olmamakla birlikte, büyük harf ile başlayıp başlamaması bu karar verilirken kullanılan başlıca kriterdir.

- **Kişi Öncesi:** Bir kişi isminden önce gelebilecek kelime veya kelime gruplarının tutulduğu dört farklı liste kullanılır. Mesleki unvanlar ("Lord, Gazi" gibi), saygı ifadeleri ("Bay, Bayan" gibi), kısaltma şeklinde mesleki unvanlar ("Dr., Prof." gibi) ve ara ifadeler ("komutanı, padişahı" gibi), bu listelerde tutulan elemanlardır.
- **Kişi Sonrası:** Bir kişi isminden sonra gelebilecek mesleki unvan ya da lakapların ("Efendi, Hatun, Han, Paşa" gibi) tutulduğu bir liste kullanılır.
- **Ülke - Devlet Sonrası:** Bir ülke ya da devlet isminden sonra gelebilecek kelime veya kelime gruplarının tutulduğu iki farklı liste kullanılır. Biri "Krallığı, Cumhuriyeti" gibi bitiş ifadeleri, diğeri "başbakanı, imparatoru" gibi ara ifadeleri bünyesinde barındırır.
- **Yer Sonrası:** Bir yer – bölge isminden sonra gelebilecek "belediye başkanı, Bölgesi, valisi" gibi bitiş ifadelerinin tutulduğu bir liste kullanılır.
- **Kurum Sonrası:** Bir kurum – kuruluş isminin sonunda yer alabilecek "Derneği, Meclisi, Kurumu" gibi bitiş ifadelerinin tutulduğu bir liste kullanılır.
- **Coğrafi Oluşum Sonrası:** Bir coğrafi oluşum isminin sonunda yer alabilecek "Gölü, Dağı, Irmağı" gibi bitiş ifadelerinin tutulduğu bir liste kullanılır. Ayrı bir kelime olarak değil, bir kelimenin sonuna eklemlenmiş şekilde karşımıza çıkabilecek "-ırmak, -dağlar" gibi ifadeler için de ilave bir liste vardır.
- **Coğrafi Olay Sonrası:** "Depremi, Yangını" gibi bitiş ifadelerinin tutulduğu bir liste kullanılır.
- **Tarihi Olay Sonrası:** "Savaşı, Devrimi, İsyanı" gibi bitiş ifadelerinin tutulduğu bir liste kullanılır.
- **Tarihi Yapı Sonrası:** "Sarayı, Köprüsü, Heykeli" gibi bitiş ifadelerinin tutulduğu bir liste kullanılır.
- **Aylar:** Yılın aylarının isimlerinin tutulduğu bir liste kullanılır.

## 3.5. Sözcüksel ve Bağlamsal Model ile Etiketleme

Sözcük birimleştirme birimindeki etiketleme işleminin aksine, sözcüksel ve bağlamsal modelde etiketleme yaparken sözcük birimler birer birer değil, n-gram yapıları şeklinde ele alınır. Bunun nedeni kullanılan kaynak listelerinde, çok sözcüklü (*multiword)* ifadelerin de bulunmasıdır. Pencere genişliği için başlangıç değeri 4 olarak belirlenmiştir ve sıfıra ulaşıncaya kadar her döngü adımında bir azaltılır. Bu yaklaşımla çok sözcüklü ifadeler ıskalanmamış ve doğru etiketlenmiş olur. Çizelge 6'da, 7 sözcük birim içeren bir cümle üzerinde n-gram tarama operasyonu için oluşturulan arama modelleri gösterilmiştir.

**Çizelge 6. N-gram tarama için oluşturulan arama modelleri (7 sözcük birim içeren cümle için)**

| N Değeri | Arama Modelleri |
|---|---|
| 4 | 1234 – 2345 – 3456 – 4567 |
| 3 | 123 – 234 – 345 – 456 – 567 |
| 2 | 12 – 23 – 34 – 45 – 56 – 67 |
| 1 | 1 – 2 – 3 – 4 – 5 – 6 – 7 |

**Çizelge 7. Sözcüksel (S) ve Bağlamsal (B) model etiketleri**

| Model | Etiket İsmi | Açıklama |
|---|---|---|
| S | LEX_TR_FN | Türkçe kişi ismi |
| S | LEX_TR_LN | Türkçe soy isim |
| S | LEX_FRGN_FN | Yabancı kişi ismi |
| S | LEX_FRGN_MN | Yabancı ikincil isim |
| S | LEX_FRGN_LN | Yabancı soy isim |
| S | LEX_CTRY | Ülke ismi |
| S | LEX_TR_CITY | Türkiye şehri ismi |
| S | LEX_TR_DIST | Türkiye ilçesi ismi |
| S | LEX_FRGN_CITY | Yabancı şehir ismi |
| S | CONJ_SWC | Büyük harfle başlayan bağlaç |
| S | NOT_LEX_SWC | Büyük harfle başlayan, sözlüksel olmayan ifade |
| B | B_PERSON | Kişi öncesi ifade |

| | | |
|---|---|---|
| B | A_PERSON | Kişi sonrası ifade |
| B | A_LOC_CTRY | Ülke – devlet sonrası ifade |
| B | A_LOC_OTH | Yer sonrası ifade |
| B | A_ORG | Kurum sonrası ifade |
| B | A_HIST_BLDG | Tarihi yapı sonrası ifade |
| B | A_HIST_EVNT | Tarihi olay sonrası ifade |
| B | A_GEO_FORM | Coğrafi oluşum sonrası ifade |
| B | A_GEO_EVNT | Coğrafi olay sonrası ifade |
| B | EW_GEO_FORM | Bitişi coğrafi oluşum belirten ifade |
| B | MONTH_NAME | Ay ismi |

Sözcüksel ve bağlamsal modeller, yararlanılan kaynak listeleri üzerinde n-gram taramalar (*lexicon lookup*) gerçekleştirerek sözcük birimleri etiketler. Kullanılan etiketler Çizelge 7 üzerinde gösterilmiştir. Şekil 2'de ise örnek bir kullanım senaryosu olarak, sırasıyla üç birimden geçerek sözcük birimleştirme ve etiketleme operasyonları tamamlanmış bir cümle gösterilmiştir. Sözcük birim etiketleri, değer atamasını yapan modele göre farklı renklerle gösterilmiştir. Etiketlemeler tamamlandığında sözcük birimlerin son hâlleri verilmiş olur ve tanıyıcı modele aktarılırlar.

## 3.6. Varlık İsimleri ve Tanıyıcı Model

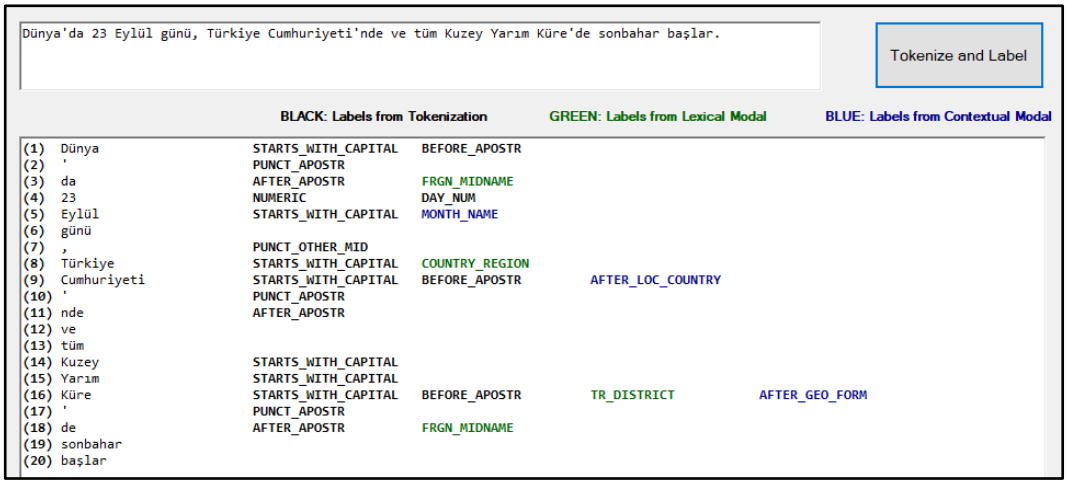
Geliştirilen VİT sistemi, tarih ve coğrafya alanlarında yazılmış ders metinleri için özelleştirildiği için, varlık ismi kavramının kapsamı da ihtiyaçlar doğrultusunda genişletilmiştir. Çizelge 8'de, sistem üzerinde tanımlanan 13 varlık ismi türü gösterilmiştir.

**Çizelge 8. Sistemde tanımlanan varlık ismi türleri**

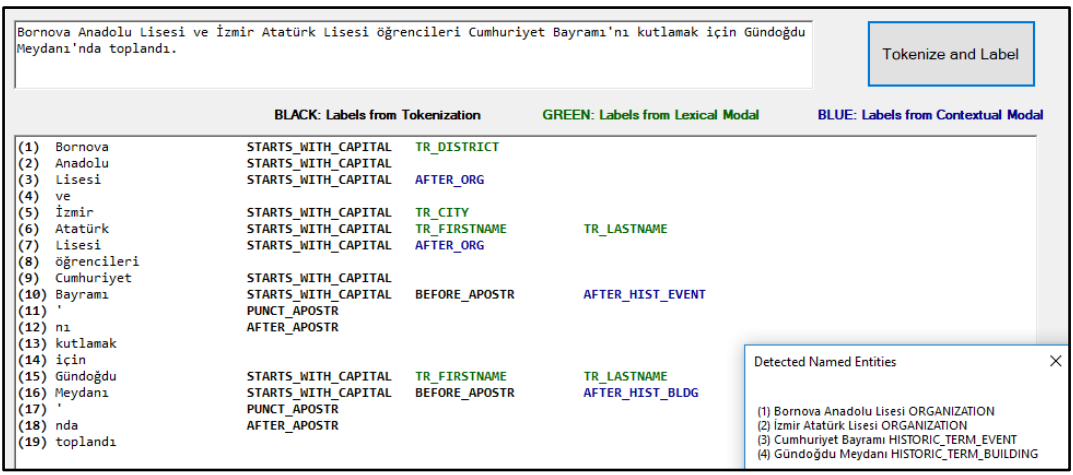
| Orijinal İsim (İngilizce) | Türkçe Açıklama |
|---|---|
| Person_Turkish | Kişi İsmi (Türk) |
| Person_Foreign | Kişi İsmi (Yabancı) |
| Location_State_Country | Yer İsmi (Ülke - Devlet) |
| Location_Other | Yer İsmi (Diğer) |
| Historic_Term_Building | Tarihi Terim (Yapı İsmi) |
| Historic_Term_Event | Tarihi Terim (Olay İsmi) |
| Geographic_Term_Formation | Coğrafi Terim (Oluşum İsmi) |
| Geographic_Term_Event | Coğrafi Terim (Olay İsmi) |
| Organization | Kurum – Kuruluş İsmi |
| Percentage | Yüzde İfadesi |
| Date | Tarih |
| Date_or_Number | Tarih veya Sayı |
| Other | Diğer |

Tanıyıcı model, varlık isimlerini tespit etmek için etiketlemesi tamamlanmış sözcük birimler üzerinde çalıştırılır. Sistem tek bir cümle ile test edilebildiği gibi, bütün hâlinde bir metin dosyası ile de çalıştırılabilir.

Şekil 3'te, sistemin "Bornova Anadolu Lisesi ve İzmir Atatürk Lisesi öğrencileri, Cumhuriyet Bayramı'nı kutlamak için Gündoğdu Meydanı'nda toplandı." cümlesi ile test edildiği örnek bir kullanım senaryosu gösterilmiştir. İşlemin sonucu olarak geriye dört adet varlık ismi döndürülmüştür.




**Şekil 2.** Sözcükbirimleştirme ve etiketleme işlemlerinden geçmiş örnek bir cümle




**Şekil 3.** Cümle üzerinde tespit edilen varlık isimlerinin gösterimine örnek bir sistem çıktısı

## 4. Araştırma Bulguları

### 4.1. Veri Kümesi ve Deney Sonuçları

Sistemin başarısı, gerçek ders metinleri üzerinde yapılan deneyler ile ölçülmüştür. Bu görev için 30 tarih ve 30 coğrafya metni seçilmiştir. Değerlendirmeler, kesinlik ve hassasiyet kriterleri ile METİN (*TEXT*) ve TÜR (*TYPE*) nitelikleri üzerinden yapılmıştır. METİN, varlık isminin sınırlarını doğru belirlemeyi; TÜR ise varlık ismi türünün doğru bulunmasını ifade etmektedir. Kıyaslama yapabilmek adına, coğrafya ve tarih alanları için deneyler ayrı olarak yapılmış; son raddede nihai sonuçlara iki deney kümesinin birleştirilmesiyle ulaşılmıştır. TÜR değeri doğru tahmin edilen varlık isimlerinin sınıflandırıldıkları tür bilgileri de sayılarak; alanlar arasındaki dağılım gözlemlenmiştir. Kesinlik, değerleri doğru tahminlerin toplam tespit edilen

**Çizelge 9. Tarih ders metinleri üzerinde yapılan deney sonuçları (İlk 5 metin gösterilmiştir)**

| BELGE İSMİ | Gerçek Vİ (#) | Bulunan Vİ (#) | Doğru METİN (#) | Doğru TÜR (#) | Bulunamayan Vİ (#) | Kesinlik METİN (%) | Kesinlik TÜR (%) | Hassasiyet METİN (%) | Hassasiyet TÜR (%) |
|---|---|---|---|---|---|---|---|---|---|
| **1. Bayezid Dönemi** | 71 | 72 | 70 | 68 | 1 | 97,22 | 94,44 | 98,59 | 95,77 |
| **1. Dünya Savaşı Öncesi Gelişmeler** | 69 | 66 | 64 | 63 | 5 | 96,97 | 95,45 | 92,75 | 91,30 |
| **1. Dünya Savaşı** | 50 | 48 | 47 | 46 | 3 | 97,92 | 95,83 | 94,00 | 92,00 |
| **1. Meşrutiyet** | 34 | 34 | 33 | 31 | 1 | 97,06 | 91,18 | 97,06 | 91,18 |
| **2. Dünya Savaşı'nın Nedenleri, Gelişimi** | 47 | 48 | 46 | 45 | 1 | 95,83 | 93,75 | 97,87 | 95,74 |
| **TOPLAM** | **1654** | **1650** | **1585** | **1529** | **69** | **96,06** | **92,67** | **95,83** | **92,44** |
| **ORTALAMA** | **55,13** | **55,00** | **52,83** | **50,97** | **2,30** | | | | |

**Çizelge 10. Coğrafya ders metinleri üzerinde yapılan deney sonuçları (İlk 5 metin gösterilmiştir)**

| BELGE İSMİ | Gerçek Vİ (#) | Bulunan Vİ (#) | Doğru METİN (#) | Doğru TÜR (#) | Bulunamayan Vİ (#) | Kesinlik METİN (%) | Kesinlik TÜR (%) | Hassasiyet METİN (%) | Hassasiyet TÜR (%) |
|---|---|---|---|---|---|---|---|---|---|
| Akarsu Havzalarımız | 34 | 33 | 32 | 31 | 2 | 96,97 | 93,94 | 94,12 | 91,18 |
| Aktif Nüfusun Ekonomik Faaliyet Gruplarına Göre Dağılımı | 14 | 14 | 14 | 14 | 0 | 100,00 | 100,00 | 100,00 | 100,00 |
| Basınç Çeşitleri ve Özellikleri | 36 | 33 | 33 | 33 | 3 | 100,00 | 100,00 | 91,67 | 91,67 |
| Başlıca Kıyı Tipleri | 29 | 29 | 28 | 27 | 1 | 96,55 | 93,10 | 96,55 | 931,10 |
| Bölgeler Coğrafyası – Akdeniz Bölgesi | 21 | 22 | 20 | 20 | 1 | 90,91 | 90,91 | 95,24 | 95,24 |
| TOPLAM | **991** | **996** | **962** | **930** | **25** | **96,59** | **93,37** | **97,07** | **94,84** |
| ORTALAMA | **33,03** | **33,20** | **32,07** | **31,00** | **0,83** | | | | |

varlık ismi sayısına bölünmesiyle; hassasiyet, değerleri doğru tahminlerin gerçek varlık ismi sayısına bölünmesiyle elde edilmiştir. Çizelge 9 ve 10, tarih ve coğrafya alanlarındaki ders metinleri için elde edilen deney sonuçlarını, Çizelge 11 ise birleştirilmiş sonuçları göstermektedir. (Çizelgelerde "varlık ismi" ifadesi "Vİ" şeklinde kısaltılmıştır.)

Deneylerde kullanılan veri kümesindeki 30 tarih metninde toplam 1654, 30 coğrafya metninde toplam 991, genel toplamda ise 60 metin için 2645 varlık ismi bulunmaktadır. Bu da bir tarih metninde ortalama 55.13, bir coğrafya metninde ortalama 33.03 ve genel ortalamada bir ders metninde 44.08 varlık ismi bulunduğunu ifade etmektedir.

Varlık ismi türlerinin metinler arasındaki dağılımını incelediğimizde, 30 tarih metninde toplam 133 Kişi

**Çizelge 11.Birleştirilmiş deney sonuçları (Toplam 60 ders metni için)**

| METİN BELGESİ | Gerçek Vİ (#) | Bulunan Vİ (#) | Doğru METİN (#) | Doğru TÜR (#) | Bulunamayan Vİ (#) |
|---|---|---|---|---|---|
| TARİH Metinleri (30) | 1654 | 1650 | 1585 | 1529 | 69 |
| Coğrafya Metinleri (30) | 991 | 996 | 962 | 930 | 25 |
| TOPLAM | **2645** | **2646** | **2547** | **2459** | **94** |
| ORTALAMA | **44,08** | **44,10** | **42,45** | **40,98** | **1,57** |

| METİN BELGESİ | Kesinlik METİN (%) | Kesinlik TÜR (%) | Hassasiyet METİN (%) | Hassasiyet TÜR (%) | F-Ölçütü METİN (%) | F-Ölçütü TÜR (%) |
|---|---|---|---|---|---|---|
| TARİH Metinleri (30) | 96,06 | 92,67 | 95,83 | 92,44 | 95,94 | 92,55 |
| Coğrafya Metinleri (30) | 96,59 | 93,37 | 97,07 | 93,84 | 96,83 | 93,61 |
| TOPLAM | **96,26** | **92,93** | **96,29** | **92,97** | **96,28** | **92,95** |

**Çizelge 12. Doğru tahmin edilen varlık ismi türlerinin tarih ve coğrafya metinlerindeki dağılımı**

| METİN BELGESİ | Kişi İsmi (Türk) | Kişi İsmi (Yabancı) | Yer İsmi (Ülke – Devlet) | Yer İsmi (Diğer) | Kurum – Kuruluş İsmi | Tarihi Terim (Yapı İsmi) |
|---|---|---|---|---|---|---|
| TARİH Metinleri (30) | 121 | 43 | 259 | 114 | 90 | 8 |
| Coğrafya Metinleri (30) | 0 | 7 | 223 | 185 | 4 | 3 |
| TOPLAM | **121** | **50** | **482** | **299** | **94** | **11** |
| ORTALAMA | **2,02** | **0,83** | **8,03** | **4,98** | **1,57** | **0,18** |

| METİN BELGESİ | Tarihi Terim (Olay İsmi) | Coğrafi Terim (Oluşum İsmi) | Coğrafi Terim (Olay İsmi) | Tarih | Tarih veya Sayı | Yüzde İfadesi | Diğer | TOPLAM |
|---|---|---|---|---|---|---|---|---|
| TARİH Metinleri (30) | 119 | 37 | 0 | 218 | 25 | 5 | 502 | **1541** |
| Coğrafya Metinleri (30) | 3 | 185 | 24 | 47 | 55 | 20 | 175 | **931** |
| TOPLAM | **122** | **222** | **24** | **265** | **80** | **25** | **677** | |
| ORTALAMA | **2,03** | **3,70** | **0,40** | **4,42** | **1,33** | **0,42** | **11,28** | |

İsmi (Türk), 48 Kişi İsmi (Yabancı), 273 Yer İsmi (Ülke – Devlet), 126 Yer İsmi (Diğer), 101 Kurum – Kuruluş İsmi, 9 Tarihi Terim (Yapı İsmi), 127 Tarihi Terim (Olay İsmi), 39 Coğrafi Terim (Oluşum İsmi), 221 Tarih, 26 Tarih veya Sayı, 5 Yüzde İfadesi ve 546 Diğer etiketli varlık ismi olduğu tespit edilmiştir. Bu metinlerde Coğrafi Terim (Olay İsmi) etiketi alan bir ifade olmadığı görülmüştür.

30 coğrafya metninde ise toplam 8 Kişi İsmi (Yabancı), 225 Yer İsmi (Ülke – Devlet), 200 Yer İsmi (Diğer), 4 Kurum – Kuruluş İsmi, 3 Tarihi Terim (Yapı İsmi), 3 Tarihi Terim (Olay İsmi), 209 Coğrafi Terim (Oluşum İsmi), 27 Coğrafi Terim (Olay İsmi), 47 Tarih, 62 Tarih veya Sayı, 20 Yüzde İfadesi ve 183 Diğer etiketli varlık ismi olduğu tespit edilmiştir. Bu metinlerde Kişi İsmi (Türk) etiketi alan bir ifade olmadığı görülmüştür.

Tarih alanı için yapılan deneylerde, METİN için %96.06 kesinlik, %95.83 hassasiyet; TÜR için %92.67 kesinlik, %92.44 hassasiyet değerlerine ulaşılmıştır. Coğrafya alanı için yapılan deneylerde, METİN için %96.59 kesinlik, %97.07 hassasiyet; TÜR için %93.37 kesinlik, %93.84 hassasiyet değerlerine ulaşılmıştır.

Sonuçlar birleştirildiğinde sistemin başarısı METİN için **%96.26** kesinlik, **%96.29** hassasiyet; TÜR için **%92.93** kesinlik, **%92.97** hassasiyet olarak ölçülmüştür. F-ölçütü değerleri ise (elde edilen kesinlik ve hassasiyet değerlerinin harmonik ortalaması alınarak) METİN için **%96.28**, TÜR için **%92.95** olarak ölçülmüştür.

Sonuçlar, coğrafya alanı için başarı oranının nispeten daha yüksek olduğunu ortaya çıkarmıştır. Ama bir tarih metninde bulunan ortalama varlık ismi sayısının, bir coğrafya metnindeki ortalama sayıdan yaklaşık 22 daha fazla olduğu da göz ardı edilmemelidir. İki alan için de METİN sonuçlarındaki doğruluğun TÜR'den yüksek olduğu (hem kesinlik hem hassasiyet değerleri için) görülmüştür. Bunun başlıca nedeni, varlık ismi sınırlarının doğru belirlenmemesi durumunda, türünün tahmin edilmesinin gerçekleştirilemez bir göreve dönüşmesidir. Kaynak listelerde yer alan çokanlamlı kelimeler ve cins isim olarak da kullanılabilen kişi isimleri, hatalı tahminlere yol açan iki diğer sebep olarak gözlemlenmiştir.

Çizelge 12, doğru tahmin edilen varlık ismi türlerinin iki alan için dağılımını göstermektedir. Diğer, Yer İsmi (Ülke – Devlet), Tarih, Kişi İsmi (Türk) ve Tarihi Terim (Olay İsmi), tarih metinlerinde en sık karşılaşılan beş varlık ismi türü olarak belirlendi. Yer İsmi (Ülke – Devlet), Yer İsmi (Diğer), Coğrafi Terim (Oluşum İsmi), Diğer ve Tarih veya Sayı ise coğrafya metinlerinde en sık karşılaşılan beş varlık ismi türü olarak belirlendi. Coğrafya metinlerinde hiç Kişi İsmi (Türk), tarih metinlerinde ise hiç Coğrafi Terim (Olay İsmi) etiketli varlık ismi olmaması dikkat çekici bir diğer sonuç olarak gözlemlendi. Yer İsmi (Ülke – Devlet), tüm deney kümesi içerisinde en homojen dağılmış varlık ismi türü olarak belirlendi.

## 4.2. Milliyet Veri Kümesi Deneyleri

Tarih ve coğrafya alanındaki ders metinleri için tasarlanmış sistemin, alan değişiminden ne düzeyde etkileneceğini görmek için, Türkçe VİT çalışmalarında sıklıkla kullanılan Milliyet veri kümesi üzerinde de deneyler yapılmıştır. Bu deneylerde, önerilen sistemin aksine varlık ismi türleri kişi, yer, kurum ve zamansal ifade ile sınırlandırılmıştır. Test kümesi ekonomi, siyaset, magazin, spor, eğlence gibi farklı kategorilere ait gazete metinleri içermektedir. Toplam 17215 kelime ve 1648 varlık ismi (623 kişi, 396 yer, 447 kurum ve 182 zamansal ifade olmak üzere) bulundurmaktadır. Çizelge 13, bu test kümesi üzerinde gerçekleştirilen deney sonuçlarını göstermektedir.

**Çizelge 13. Milliyet Veri Kümesi Deney Sonuçları**

| | Kesinlik (%) | Hassasiyet (%) | F-Ölçütü (%) |
|---|---|---|---|
| **METİN** | 91.46 | 94.17 | 92.80 |
| **TÜR** | 84.38 | 86.89 | 85.62 |

Alan değişikliğinin sistem başarısını bir miktar düşürdüğü gözlemlenmiştir. Sistemin haber metinleri üzerindeki başarısı f-ölçütü ile METİN için **%92.80**, TÜR için **%85.62** olarak ölçülmüştür. METİN sonuçlarındaki başarının TÜR'den (hem kesinlik, hem hassasiyet değerleri için), hassasiyet sonuçlarındaki başarının da kesinlikten yüksek olduğu belirlenmiştir. Varlık ismi türlerine göre başarı incelendiğinde ise, hem METİN hem TÜR değerinin doğru tespit edilme oranı kişi isimleri için %**78.01**, yer isimleri için **%93.18**, kurum isimleri için **%79.64,** zamansal ifadeler için **%98.35** olarak ölçülmüştür.

### A. 4.3. Karşılaşılan Güçlükler

Sistemin geliştirilmesi süresince, Türkçe dilinin yapısından ya da girdi metni dosyalarında ihlallerden kaynaklı bazı güçlük ve kısıtlamalar ile karşılaşıldı.

Şekil 3'te verilen kullanım senaryosu örneğinde, "Bornova, İzmir, Atatürk, Gündoğdu" sözcük birimleri sözlüksel ifadeler olarak etiketlenmiştir ve farklı bir metin içeriğinde kendi başlarına da birer varlık ismi ifade edebilirler. Sistem bu gibi durumlarda kapsayıcı terimi dikkate alacak şekilde tasarlanmıştır. Şekilde görüldüğü üzere, bu terimler isabetli bir şekilde, daha uzun varlık isimlerinin parçası olarak görülmüşlerdir.

Türkçe kişi isimlerini tespit etmek için geniş bir sözlük yapısı kullanmak, hassasiyet kriterini olumlu etkilese de kesinlik kriterine olumsuz etkisi olması muhtemeldir. Bunun sebebi, bazı Türkçe kişi isimlerinin, ders metinlerinde sıkça karşılaşılan cins

isimleri işaret edebiliyor oluşudur. Bu duruma "Savaş, Barış, Nehir, Irmak" gibi kelimeler örnek olarak verilebilir. Bu ifadeler cümlenin başında olduğunda, komşu kelimelerin kontrol edilmesi büyük oranda sorunu çözmektedir; ancak tek başına yeterli olmadığı durumlar görülebilir. Örneğin, CONJ_SWC listesi cümlenin bir bağlaç ile başlayıp, bir varlık ismi ile devam ettiği durumlarda hataların önüne geçmek için oldukça önemli bir kaynaktır.

"Sultan, Şah" gibi bağlamsal modelde kullanılan bazı ifadeler, bir kişi isminden önce de sonra da karşımıza çıkabilir; hatta bazen "Kanuni Sultan Süleyman" gibi ifadelerde iki durum aynı anda gerçekleşebilir. Sistem iyileştirmeler yapılmadan önce, bu gibi durumlarda çıktı olarak iki farklı varlık ismi ("Kanuni Sultan" ve "Sultan Süleyman" şeklinde) vermeye meyilliyken; bu durum düzeltilerek yarım ifadelerin birleştirilip tek ve doğru varlık isminin verilmesi sağlanmıştır.
Başlıklar, varlık ismi barındırma ihtimali yüksek olduğu için sistem tarafından değerlendirmeye alınır. Ama genel kullanımda, başlık metinlerindeki bütün kelimelerin (eğer bağlaç değillerse) ilk harfleri, özel isim olmasa dahi büyük yazılmaktadır. Bu durum hatalı varlık ismi tespitlerine yol açabilmektedir. Bunu önlemek adına sisteme, üzerinde çalıştığı metnin bir başlık mı yoksa bir cümle mi olduğu bilgisi verilerek; eğer başlık ise tanıyıcı modelde "Diğer" etiketi için yapılan kontroller çalıştırılmamaktadır. Yine de, kesme işareti kontrolleri yapılmaya devam edildiği için "Diğer" etiketli bir varlık isminin tespit edilmesi ihtimali ortadan kalkmış değildir.

Genelde literatürde "Kişi İsmi" şeklinde kullanılan varlık ismi türünün "Kişi İsmi (Türk)" ve "Kişi İsmi (Yabancı)" şeklinde ikiye ayrılmasının bazı durumlarda, iki etiket birleştirilse gerçekleşmeyecek hatalı TÜR sonuçlarına yol açtığı gözlemlendi. Bunun sebeplerinden en önemlisi "Musa, Enver, Zeynel, Süleyman" gibi Türkçede kullanılan bazı kişi adlarının, Arap ülkelerinde de kullanılıyor olmasıdır. Yine de deney sonuçları ve yapılan ayrımın gelecekte sisteme sağlayabileceği faydalar baz alındığında, bu durumdan kaynaklı performans kaybı kabul edilebilir düzeydedir.

Girdi olarak sisteme verilen metinlerdeki önemli noktalama işareti eksiklikleri (kesme işareti ve virgül gibi) ve yazım hataları sistem başarısını aşağıya çekecektir. Aynı zamanda bulunan varlık isimlerinin kalitesini düşürecek, "Diğer" etiketi alan varlık ismi sayısını arttıracaktır. Bu nedenle metin dosyalarının sistem kullanımına sunulmadan önce yazım denetiminden geçirilmesi tavsiye edilir.

## 5. Sonuçlar ve Öneriler

Bu çalışmada, kapsamı tarih ve coğrafya alanları olarak belirlenen Türkçe ders metinleri için kural tabanlı bir VİT modeli geliştirilmiştir. Sistem girdi olarak bir metin dosyası alıp, metin içeriğini tarayarak varlık isimlerini tespit edecek ve bulguları çıktı olarak sunacak şekilde tasarlanmıştır. Geliştirilen model ve modelin kullanımı için oluşturulan sözlüksel kaynaklar, Dokuz Eylül Üniversitesi Doğal Dil İşleme (NLP) sunucusu bünyesinde tutulmaktadır.

Sistemin başarısı, Tarih ve Coğrafya ders metinleri üzerinde yapılan deneyler ile ölçülmüştür. 30 tarih ve 30 coğrafya metni rastgele seçilmiş, değerlendirmeler kesinlik ve hassasiyet kriterleri ile METİN ve TÜR nitelikleri üzerinden yapılmıştır. Sonuç olarak sistemin başarısı METİN için %96.26 kesinlik, %96.29 hassasiyet; TÜR için %92.93 kesinlik, %92.97 hassasiyet olarak ölçülmüştür.

### 5.1. Çalışmanın Eğitimsel Değeri

Çalışma doğal dil işleme, bilgi çıkarımı, metin madenciliği ve bilişimsel dilbilim alanlarının kapsamında olmakla beraber; bilgisayar destekli bir eğitim yazılımı olarak değerlendirmek de mümkündür.

Araştırmaya başlarken konulan birincil hedefler değerlendirildiğinde, deneylerden büyük oranda tatmin edici sonuçlar elde edilmiştir. Bu durum, geliştirilen VİT modelinin, uzun vadeli eğitimsel hedef olan tarih ve coğrafya alanlarında kullanılabilecek nitelikli ve esnek yapılı terimler sözlükleri elde etmek için uygun bir yardımcı araç olabileceğini ortaya koymuştur.

Çalışmada 13 farklı varlık ismi türü tanımlanmıştır. Bu sayede sistem, oldukça geniş bir "tarihi terim" ve "coğrafi terim" sınıflandırması yerine, terimler için daha anlamlı bir tasnif modeli önermektedir. Terimler sözlüğü yapıları, soru niteliği taşıyan varlık isimlerinden meydana geleceği için; bu yapıların sınav hazırlama süreçlerine yardımcı olabileceği öngörülmektedir.

## 5.2. İyileştirme Olanakları

Sonuçların kalitesini arttırmak adına, "Diğer" etiketli varlık isimlerinin sayısını azaltmak hedeflenebilir. Bunun için de ilave varlık ismi türleri tanımlanabilir. Örneğin, "Diğer" etiketi alan tarih metinlerindeki varlık isimlerinin büyük bir bölümünün millet, milliyet anlamı taşıyan ifadeler olduğu görülmüştür. Bu ifadelerin yeni ve daha anlamlı bir türün kapsamına alınmasının üzerinde durulabilir. Sözlüksel kaynaklar, antik çağ yer ve kişi isimlerini de barındıracak şekilde genişletilebilir. Noktalama işareti eksikliğinden kaynaklı olumsuz etkiyi azaltmak için, bir yazım denetimi biriminin sisteme dahil edilmesi düşünülebilir.

## Teşekkür

Bu makale Dokuz Eylül Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimi (DEÜBAP) tarafından 2018.KB.FEN.015 numarasıyla desteklenen proje çalışması kapsamında hazırlanmıştır.

## Kaynakça

[1] Jurafsky, D., Martin, J.H. "*Speech and language processing (2nd Edition)"*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. (2009)

[2] Grishman, A., Sundheim, B. "Message Understanding Conference-6: a brief history". In Proceedings of the 16th conference on Computational linguistics - Volume 1 (COLING '96), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 466-471. (1996)

[3] Cucerzan, S., Yarowsky, D. "Language independent named entity recognition combining morphological and contextual evidence". Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. New Brunswick, NJ: Association for Computational Linguistics. (1999)

[4] Alfonseca, E., Manandhar S. "An unsupervised method for general named entity recognition and automated concept discovery". In 1st International Conference on General WordNet. (2002)

[5] Tür, G., Hakkani-Tür G., Oflazer K. "A statistical information extraction system for Turkish". Natural Language Engineering, vol. 9 (2), pp. 181-210 (2003)

[6] Sang, E., Meulder F. "Introduction to the CoNLL-2003 shared task: language-independent named entity recognition". In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4 (CONLL '03), Vol. 4. Association for Computational Linguistics, Stroudsburg, PA, USA, 142-147 (2003)

[7] Wentland, W., Knopp, J., Silberer, C., Hartung, M. "Building a multilingual lexical resource for named entity disambiguation, translation and transliteration". in Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco. (2008)

[8] Küçük, D., Yazıcı, A. "Rule-based named entity recognition from Turkish texts". In Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, Trabzon, Turkey. pages 456–460. (2009)

[9] Küçük, D., Yazıcı, A. "Named entity recognition experiments on Turkish texts". In Proceedings of the 8th International Conference on Flexible Query Answering Systems, FQAS '09, pages 524–535, Berlin, Heidelberg. Springer-Verlag. (2009)

[10] Tatar, S., Çiçekli, İ. "Automatic rule learning exploiting morphological features for named entity recognition in Turkish". Journal of Information Science, 37 (2), 137-151. (2011)

[11] Küçük, D., Yazıcı, A. "A hybrid named entity recognizer for Turkish with applications to different text genres". In: Gelenbe E., Lent R., Sakellari G., Sacan A., Toroslu H., Yazici A. (eds) Computer and Information Sciences. Lecture Notes in Electrical Engineering, vol 62. Springer, Dordrecht. (2012)

[12] Şeker, G. A., Eryiğit, G. "Initial explorations on using CRFs for Turkish named entity recognition". In Proceedings of COLING 2012, Mumbai, India. (2012)

[13] Küçük, D., Jacquet, G., Steinberger, R. "Named entity recognition on Turkish tweets". In: Language Resources and Evaluation Conference. (2014)

[14] Küçük, D., Küçük, D., Arıcı, N. "A named entity recognition dataset for Turkish". In: 24th Signal Processing and Communications Applications Conference (SIU), Zonguldak, Turkey. (2016)

[15] Şeker, G., Eryiğit, G. "State of the art in Turkish named entity recognition". https://pdfs.semanticscholar.org/7e7f/ed9d21a3e3a36c4eb3c7df1ee8116e8ec2ce.pdf (2016)

[16] Ertopçu, B., Kanburoğlu, A., Topsakal, O., Açıkgöz, O., Gürkan, A., Özenç, B., Çam, İ., Avar, B., Ercan, G.,

Yıldız, O. “A new approach for named entity recognition”. In: International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey. (2017)

[17] H. B. Şahin, C. Tirkaz, E. Yildiz, M. T. Eren, and O. Sonmez, “Automatically annotated turkish corpus for named entity recognition and text categorization using large-scale gazetteers”.arXiv preprint arXiv:1702.02363, (2017)

[18] Güneş, A., Tantuğ, A. C., “Turkish named entity recognition with deep learning”. 26th Signal Processing and Communications Applications Conference (SIU). doi:10.1109/siu.2018.8404500 (2018)

[19] Güngor, O., Üsküdarlı, S., Güngor, T., “Recurrent neural networks for Turkish named entity recognition”. 26th Signal Processing and Communications Applications Conference (SIU). doi:10.1109/siu.2018.8404788 (2018)